DS 5110: Big Data Systems | Final Project
State of Virginia Traffic Reliability – MAP21
Christian Schroeder (dbn5eu), Timothy Tyree (twt6xy), Colin Warner (ynq9ya)

i. Abstract

a. Introduction

This study used Virginia specific data for a set of independent variables (Year, Period, Precipitation Rate, Volume/Capacity Ratio, Hourly Volume, Presence of Safety Service Protocol, Crashes, Weather Events, Number of Lanes, County, Terrain, Urban Designation, Road Direction, Intersection, Segment Order and Truck Percentage) to predict if a MAP-21 reporting segment was reliable.

Primary Objective: Use Virginia highway traffic data from 2017-2020 to accurately predict the reliability of the state's traffic projections. If that model is found, we can use the state's forecasted metrics through 2024 to classify *future* unreliable highway segments.

Context: In 2012, President Obama signed into law the Moving Ahead for Progress in the 21st Century Act (MAP-21). Among other initiatives, this act transforms the process used for allocating funds towards the improvement of highway, transit, bike, and pedestrian programs - allowing a programmatic framework to inform whether or not a road is in need of transformation.

As part of an ongoing project at the Virginia Department of Transportation (VDOT), our team has been asked to explore additional classification models to predict if a MAP-21 reporting segment is reliable.

b. Overview of Process

Below we provide a quick description of each stage in our process. Please see specified sections in parentheses for a more thorough description of each stage.

Data Import and Preprocessing (section ii.a.): Write Preprocessing Class that (a) handles the import of packages and initializes the Spark Session, and then (b) reads, combines, and transforms data from 12 separate csv files to a workable format.

Data Splitting (section ii.b.): Split combined data into 'actual' and 'forecasted' data sets prior to Exploratory Data Analysis. Split the 'actual' data into train (90%) and test (10%) segments. Forecasted data is held out to use for classifying future unreliable segments.

Exploratory Data Analysis (section ii.c.): Evaluate distributions of numeric variables to determine necessary transformations. Three numeric variables benefit from log transformations. Also determined it necessary to drop geographic categorical variables due to

certain instances of these variables not having examples of unreliable segments. Finally, we constructed a visualization to display highway segment data on a map of Virginia.

Model Construction (section iii.a.): Built pipeline for treating categorical variables as factors using StringIndexer and OneHotEncoder, as well as adding independent variables for modeling to a features vector. Constructed Logistic Regression, Random Forest, and Decision Tree models using cross validation.

Model Evaluation (section iii.b.): After evaluating our three model types on the basis of AUROC and accuracy, as well as the true positive, true negative, false positive, and false negative rates, we decided that the Decision Tree model with a max depth of 10, max bins of 20, and threshold of 0.5.

c. Summary of Finding

The Decision Tree model projected on average 136 highway segments to be unreliable each year from 2021-2024 (section iv.a.):

Our best model identified on average 136 highway segments as unreliable during either the morning, midday, evening, or weekend traffic periods for the forecasted years of 2021-2024. We believe the accuracy of our model is reliable and we recommend VDOT prioritize these projected unreliable highway segments when allocating funds for highway improvements.

Tradeoffs based on context of the problem (section iv.b.):

We had the opportunity to select a lower threshold for our Decision Tree model that would increase the AUROC. However, we believe that VDOT would care a great deal about limiting the number of false positives to avoid spending money on a highway segment that is actually reliable. This assumption led us to trade a model with a slightly higher AUROC for a model with a significantly lower number of False Positives.

Future Work (section iv.c.):

We believe that the Decision Tree model we identified in this project enables the Department of Transportation the opportunity for additional work on this topic. Two areas that we would explore further if we had time are (1) evaluating the methodology used for the Forecasted data set, and (2) developing methods to advise the state on which enhancements would make specific unreliable segments more reliable.

ii. Data and Methods

a. Data Import and Preprocessing

The first challenge faced was the formatting of the data. Our team was provided 12 datasets, each representing one or more different predictors. The main issue in combining the data came with discrepancies in the grouping of data, whether it be by only highway segment (TMC), both TMC and Year, or by TMC, Year, and Period. To tackle this, we created the Preprocessing class to read and format the data. Below is an outline of the steps included in a readAndCombineData() function that is part of the preprocessing class:

1. Create a dictionary of directories with the directory name as key (ex. TMC/) and empty lists as values. This will hold dataframes that can be joined on shared unique identifiers.
2. Gets the full path to the input directories and uses a formatted string to get get the other directories in a loop. (looping through the directory name keys)
3. Creates a nested list of lists that define the type of joins each directory will be performing. Ordered the same as the directories.
4. Joins all data as follows;
    - a) Outer loop through each directory
    - b) Inner loop through the files in each directory and read the file into a Spark dataframe.
    - c) Append the dataframe to the values list within its respective directory (key/outer loop)
    - d) Get out of the inner loop, pop the last data frame out of this list, and save it to a temporary variable. This will be the df that starts joining on each directories respective join identifiers.
    - e) Join dataframes within each directory into one. Results in 3 dataframes after starting with 12. The logic is similar to sorting algorithms. Within another inner loop Start with the dataframe that was popped, set a temp df as that df, create a joined df with the temp df and the current loops df, on the columns specified within the current iterations index location of the join list, then set the temp df as the joined df, and set the start df back to temp. This will successfully join each dataframe within the list on their respective identifiers without repeating or missing dataframes.
    - f) Outside the previous loop, append the start df (which is now the full joined df) to the end of the list, and drop every other df in the list.
    - g) As a sanity check, loop through all columns in the joined df and drop any that may be duplicate.
5. Sequentially join the final three dfs into one, making sure to join on the df that had more previous identifiers so there's no data loss.
6. Create trainable and forecasted datasets by filtering on year (trainable < 2021), (forecasted > 2020)
7. Save the data.

b. Data Splitting

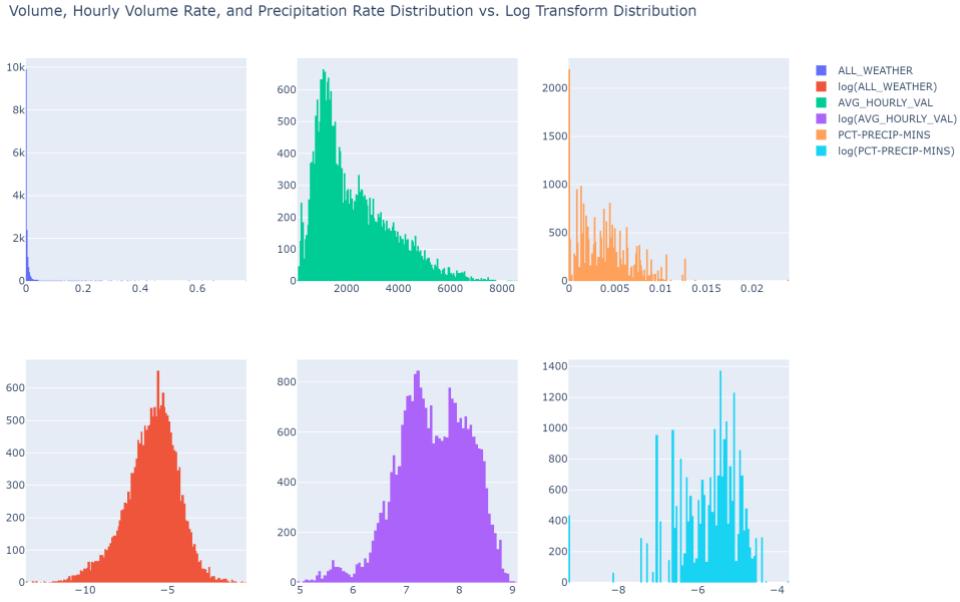
As briefly mentioned in the outline of our Preprocessing class, the single resulting dataframe from preprocessing contains ‘actual’ data from 2017-2020, as well as VDOT’s ‘forecasted’ data through 2024. This was split into the trainable and forecasted dataframes, with 54,624 and 27,312 observations, respectively.

We then took the trainable data set and split that further into both train and test sets. We elected to use a 90%/10% split given our relatively small data set after holding out the forecasted data. We used these train and test sets to evaluate several model types and determine whether there is a model that can accurately identify unreliable highway segments.

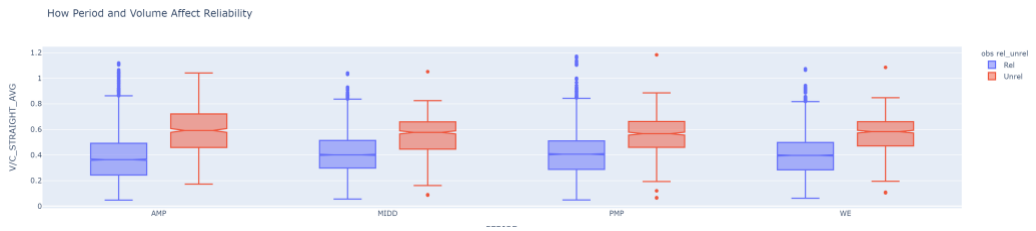
The forecasted data held the identical shape as the trainable data, with 22 columns and 27,312 observations. We used the best performing model to identify unreliable highway segments in the forecasted data.

c. Exploratory Data Analysis

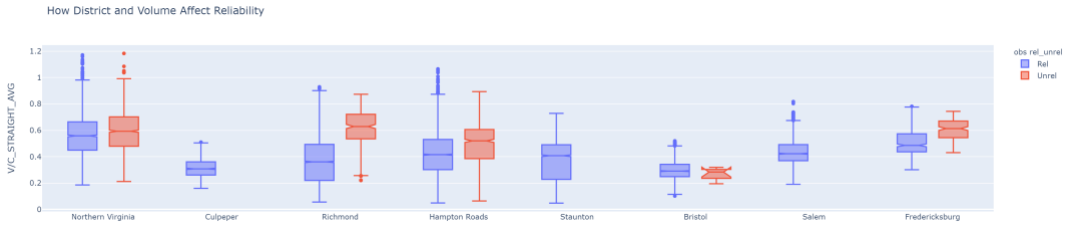
Before modeling, we needed to fully understand our data and how it behaved. Our first step of ETA was to evaluate the distribution of numerical variables to determine if any transformations were necessary. Upon this evaluation, it was determined that three numeric variables would benefit from a log transformation. These numeric variables included Weather, Precipitation, and Hourly Traffic Volume.



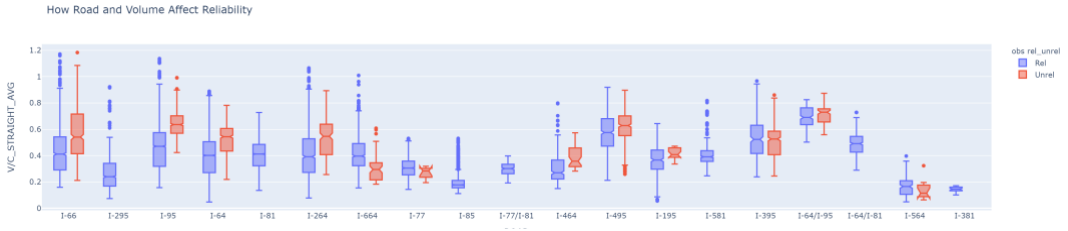
After reviewing the numerical columns, we inspected the categorical predictors. One of those predictors being Period, the time of day. We found that the peak AM and PM periods (AMP and PMP) had more unreliable instances than the others. This was expected because the volume of traffic would be greater at those times.



We also investigated the geographic variable, District, the area of the state the segment is within. We found that there were three districts that contained only reliable segments; Culpeper, Staunton, and Salem.

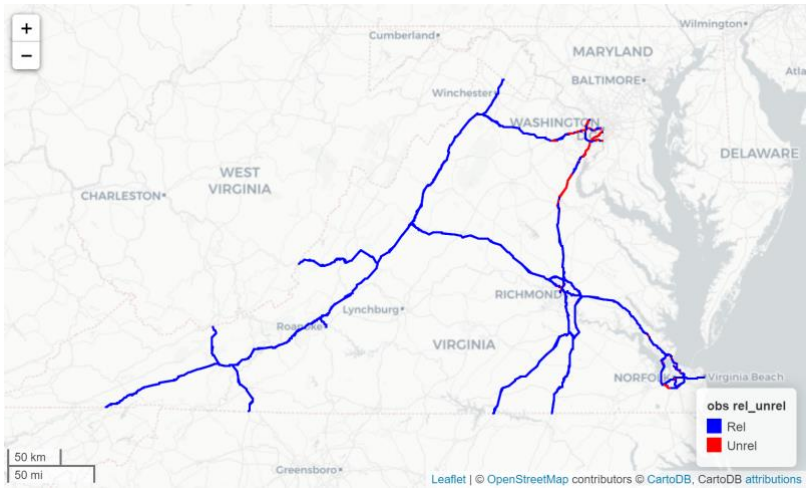


A similar situation was seen in the Road variable with several interstates showing no unreliability, including I-81 and I-77.

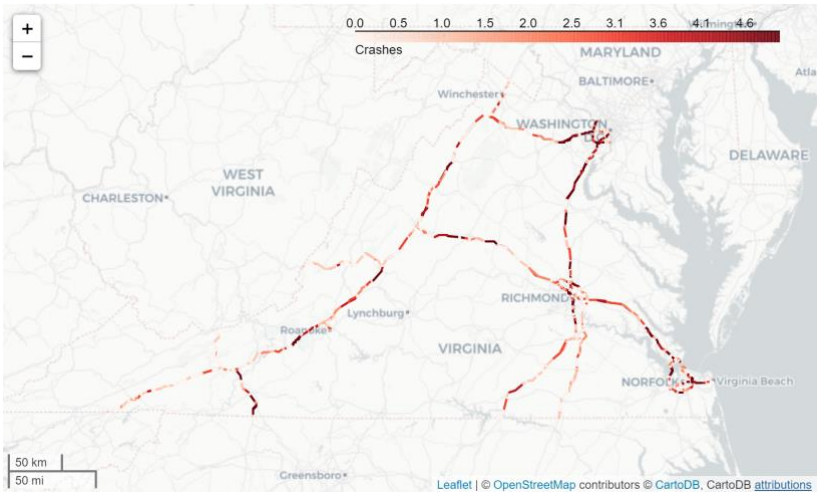


Ultimately, we believed the best way to move forward was to omit both variables, Road and District, from the final trainable dataset. This was done to allow a higher possibility of segments in those areas or on those roads being flagged as unreliable.

To help us get a better understanding of the spatial relations of our data, a Mapper class was created to generate interactive maps of the highway segments. When looking at the distribution of reliable and unreliable segments, we found that most of the unreliable roads were in the Norfalk/Virginia Beach, Richmond, and Northern Virginia areas. This was not surprising given their populations relative to the rest of the state.



This distribution of unreliability was expectedly like the distribution of average hourly volume. Unexpectedly, the distribution of crashes across the state did not seem to correlate with population centers. Rather the crashes were spread generally evenly across the state, with possible hotspots around Northern Virginia and Norfolk.



The noticeable presence of crashes along I-81 did not seem to impact the reliability of the interstate, as no segments were marked as such. The lack of expected influence on reliability could have been due to the almost random spread of crashes, or it's possible that crashes do not affect overall reliability and traffic time in the long run.

iii. Results

a. Model Construction

PySpark's Pipeline functionality was used to streamline our data preparation for modeling. The pipeline included stages for string indexing and one hot encoding for categorical variables, and the final vector assembling to create the features column. The features column consisted of vectors containing all independent variables that would be input into the models.

Logistic Regression:

We started with a simple logistic regression model with no tuning to get a baseline understanding of the ability to classify the segments. This preliminary model performed well, producing an accuracy of 0.91 and an AUROC of 0.74. However, there were a lot of false positive and negative predictions.

To improve on the reliability of the model we felt it was necessary to implement cross validation. We created a parameter grid with possible values for lambda, alpha, and max iterations. Then each possible model was trained on the training data with 10-fold cross validation. The best model, chosen by the CrossValidator() object based on AUROC, had a lambda of 0.1, alpha of 0.2, and max iterations of 10. This model also had an accuracy of 0.91 but an AUROC of 0.66.

We further investigated the model by testing varying thresholds to see how the AUROC could possibly be improved. We found the highest AUROC of 0.87 with a threshold of 0.1, but that came with a steep increase in false positives.

Threshold	TP	TN	FP	FN	Accuracy	AUROC
0.1	0.934	0.811	0.189	0.066	0.82411	0.8727
0.2	0.737	0.904	0.096	0.263	0.88638	0.82049
0.3	0.554	0.943	0.057	0.446	0.90168	0.74812
0.4	0.457	0.965	0.035	0.543	0.91187	0.71108
0.5	0.343	0.978	0.022	0.657	0.91114	0.66029
0.6	0.211	0.988	0.012	0.789	0.90605	0.59943
0.7	0.093	0.996	0.004	0.907	0.90095	0.54468
0.8	0.031	0.999	0.001	0.969	0.89694	0.51496
0.9	0.003	1.0	0.0	0.997	0.89512	0.50173

As it would be in the best interest of VDOT to not waste money on already reliable segments, we believed the best option was a threshold of 0.4, where false positives did not increase severely, but the AUROC would go up to 0.71.

Random Forest:

The second model we trained was random forest. At each tree split, a random subset of the predictors is chosen. This method gives less strong variables more of a chance to have an influence. We used cross validation to select the best model given the tuning parameters of max depth, max bins, and number of trees. The best performing Random Forest model against the training data had a max depth of 10, max bins of 10, and number of trees at 5. This model, when used to predict the test data, had an accuracy of 0.90 and AUROC of 0.54.

This model performed worse on the test data than the logistic regression model, so we tested additional threshold values to try and improve the AUROC. The highest AUROC of 0.82 was found at a threshold of 0.2, which came with a very large drop in accuracy to 0.70 and a high false positive rate.

Threshold	TP	TN	FP	FN	Accuracy	AUROC
0.2	0.7889	0.8596	0.1404	0.2111	0.85215	0.82426
0.1	0.9827	0.6227	0.3773	0.0173	0.6606	0.8027
0.3	0.5571	0.9573	0.0427	0.4429	0.91515	0.75718
0.4	0.1972	0.9923	0.0077	0.8028	0.90859	0.59475
0.5	0.083	0.9996	4.0E-4	0.917	0.90313	0.54132
0.6	0.0415	1.0	0.0	0.9585	0.89913	0.52076
0.7	0.0173	1.0	0.0	0.9827	0.89658	0.50865
0.8	0.0	1.0	0.0	1.0	0.89476	0.5
0.9	0.0	1.0	0.0	1.0	0.89476	0.5

We believed the best threshold for this model was also 0.4, with an AUROC of 0.59 and accuracy of 0.91 because of the relatively low false positive rate.

Decision Tree:

After seeing the performance of the random forest model, we wondered how a simple decision tree model would perform on the same data. It was possible that certain predictors played a bigger role in segments being unreliable or not, and we were suppressing that influence. With the tuning parameters max depth and max bins, the cross validation found the best model to have a max depth of 10 and max bins of 20. When used to predict the test data, this model had an accuracy of 0.94, AUROC of 0.84, and only 68 false positives.

This model performed better than the previous two on all three metrics. We then tested the varying thresholds to see how it would affect those values. The best AUROC was at a threshold of 0.1 and came with a higher false positive rate.

Threshold	TP	TN	FP	FN	Accuracy	AUROC
0.1	0.803	0.901	0.099	0.197	0.89039	0.85173
0.2	0.699	0.974	0.026	0.301	0.94465	0.83625
0.3	0.692	0.978	0.022	0.308	0.94756	0.83483
0.4	0.689	0.978	0.022	0.311	0.94756	0.8333
0.5	0.689	0.978	0.022	0.311	0.94756	0.8333
0.6	0.498	0.994	0.006	0.502	0.9421	0.74629
0.7	0.484	0.996	0.004	0.516	0.9421	0.74018
0.8	0.474	0.996	0.004	0.526	0.94101	0.73499
0.9	0.415	0.998	0.002	0.585	0.937	0.7068

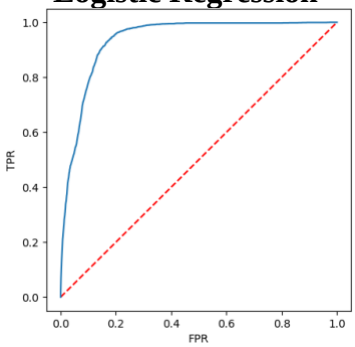
We found that the best thresholds were between 0.5 and 0.2, where the AUROC, accuracy, and false positives remained relatively unchanged. From this we decided changing the threshold from 0.5 did not add any benefit.

b. Model Evaluation

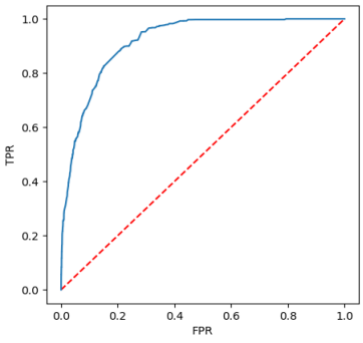
When evaluating the performances of the chosen models on the test data, we looked at the counts of true positive, true negative, false positive, and false negative predictions, as well as the accuracy and AUROC. The metrics we found the most important were accuracy, AUROC, and the false positive predictions. The ROC curves were also plotted and evaluated.

Model	Threshold	TP	TN	FP	FN	Accuracy	AUROC
LR	0.4	0.457	0.965	0.035	0.543	0.91187	0.71108
RF	0.4	0.197	0.992	0.008	0.803	0.90859	0.59475
DT	0.5	0.689	0.978	0.022	0.311	0.94756	0.8333

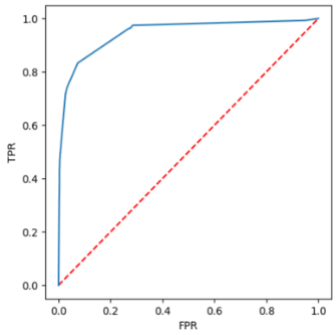
Logistic Regression



Random Forest



Decision Tree



Although the decision tree model did not have the lowest false positive rate, it did have noticeably better accuracy and AUROC values. the We decided to move forward with the decision tree model due to its ability to perform well in Accuracy and AUROC while maintaining a reasonable false positive rate.

iv. Conclusions

a. Decision Tree projected average of 136 highway segments to be unreliable each year from 2021-2024

The primary goal of our project was to identify future unreliable highway segments for VDOT. According to the state, an unreliable segment must only be classified by a model as unreliable in either the morning, midday, evening, or weekend to be considered unreliable. Our decision tree model has identified on average 136 segments that are projected to be unreliable each year from 2021-2024. We believe these segments should be prioritized when allocating MAP-21 funds to highway improvements.

Year	sum(prediction)	count
2021	0.0	1571
2021	1.0	34
2021	2.0	12
2021	3.0	24
2021	4.0	66
2022	0.0	1573
2022	1.0	32
2022	2.0	11
2022	3.0	25
2022	4.0	66
2023	0.0	1573
2023	1.0	34
2023	2.0	9
2023	3.0	25
2023	4.0	66
2024	0.0	1569
2024	1.0	37
2024	2.0	6
2024	3.0	25
2024	4.0	66

→ 136 total unreliable

→ 134 total unreliable

→ 134 total unreliable

→ 138 total unreliable

b. Tradeoffs based on context of the problem

When multiple model types provide sufficient accuracy in predicting unreliable highway segments, it becomes important to understand the context of the problem in making a decision. For example, one assumption that we made about this problem was that it was essential VDOT limits false positives. If we are dealing with distributing limited funds towards the improvement of highway infrastructure, we want to ensure that we are not accidentally recommending that funds be used on a segment of highway that is reliable.

In our instance, this made a big difference in our final model selection. After deciding on Decision Tree as the best model type, we tested the Decision Tree across a range of threshold values. Lowering the threshold actually increased the AUROC of our model. However, based on our assumption of needing to limit the False Positives, we chose to stick with our original threshold of .5.

c. Future work

We were provided a data set that forecasted each of our predictors for the years 2021-2024. While we understand from the Virginia Department of Transportation that they generally trust the forecasted data, it would be of interest to our group to understand the ins and outs of developing that data set. One observation we made is the general lack of change in data from one year to another for any given traffic segment. This makes sense given the mostly static nature of a highway segment, but we do wonder if there are other variables that may influence a highway segment to be more unreliable one year compared to another.

Our work to this point has focused on simply identifying unreliable segments for the Virginia Department of Transportation. Our models have not yet dealt with the actions the VDOT should take to make these segments of Virginia highways more reliable in the future. In order to do this, we could start by focusing more attention on the coefficients of the variables in the model – which factors are most influential in determining unreliability, and can those factors be manipulated to enhance reliability?